data\_refining.R

USER

Fri Nov 16 01:11:19 2018

#  
# PROJECT\_1. 보건.복지 - 사망원인통계  
#  
# COPYRIGHT (c) 2018 AJOU University  
# Author : Pae, Jaehoon  
# History : 2018/10/12  
#

#install.packages(c("readxl", "dplyr"))  
library(readxl)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

#2016년 사망원인 데이터 읽어오기  
annual <- read.csv("사망원인\_2016.csv", header=FALSE)  
  
#분석에 사용할 항목만 데이터 선택하기  
death <- annual[,c(2, 4, 5, 7, 9:14, 20)]  
  
#column 이름 부여하기  
names(death) <- c("신고날짜(월)", "사망자주소", "사망자성별", "사망일자(월)", "사망시간", "사망연령",   
 "사망장소", "사망자직업", "혼인상태", "교육정도", "사망원인")  
  
#사망자 주소 전처리  
death$사망자주소 <- factor(death$사망자주소,   
 levels=c(11, 21, 22, 23, 24, 25, 26, 29,  
 31, 32, 33, 34, 35, 36, 37, 38, 39),  
 labels=c("서울", "부산", "대구", "인천", "광주", "대전", "울산", "세종",   
 "경기", "강원", "충북", "충남", "전북", "전남", "경북", "경남", "제주"))  
#사망장소 전처리  
death$사망장소 <- factor(death$사망장소,  
 levels=c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 99),  
 labels=c("주택", "의료기관", "사회복지시설", "공공시설",   
 "도로", "상업.서비스시설", "산업장", "농장",   
 "병원 이송 중", "기타", "미상"))  
#혼인상태 전처리  
death$혼인상태 <- factor(death$혼인상태,  
 levels=c(1, 2, 3, 4, 9),   
 labels=c("미혼", "배우자있음", "이혼", "사별", "미상"))  
  
#교육정도 전처리  
death$교육정도 <- factor(death$교육정도,   
 levels=c(1, 2, 3, 4, 6, 7, 9),  
 labels=c("불취학", "초등학교", "중학교",  
 "고등학교", "대학교", "대학원이상", "미상"))  
  
#직업 코드 시트  
code\_job <- read\_excel("2016년 사망(원인) 통계 설계서 및 코드집 \_ 공공용.xlsx", col\_names = T, sheet = 3)  
  
job <- code\_job[c(2:11),c(1, 2)] #직업코드  
job[, c(1)] <- sapply(job[,c(1)], as.numeric) #직업코드 정수화  
  
#사망자직업 전처리  
death$사망자직업 <- factor(death$사망자직업,  
 levels=job$`2008~2016년`,  
 labels=job$X\_\_1)  
  
#사망원인 코드 시트  
code\_reason <- read\_excel("2016년 사망(원인) 통계 설계서 및 코드집 \_ 공공용.xlsx", col\_names = T, sheet = 5)  
names(code\_reason) <- c("가", "나", "다")  
code\_reason <- subset(code\_reason, !is.na(가)) #na제거  
  
reason <- code\_reason[c(2:57), c(1, 2)] #사망원인 코드  
reason[, c(1)] <- sapply(reason[,c(1)], as.numeric) #코드 정수화  
#사망원인 전처리  
death$사망원인 <- factor(death$사망원인, levels=reason$가, labels=reason$나)  
  
summary(death) #요약

## 신고날짜(월) 사망자주소 사망자성별 사망일자(월)   
## Min. : 1.000 경기 : 55215 Min. :1.000 Min. : 1.000   
## 1st Qu.: 3.000 서울 : 43540 1st Qu.:1.000 1st Qu.: 3.000   
## Median : 6.000 경남 : 21193 Median :1.000 Median : 6.000   
## Mean : 6.415 부산 : 21074 Mean :1.457 Mean : 6.464   
## 3rd Qu.: 9.000 경북 : 20978 3rd Qu.:2.000 3rd Qu.:10.000   
## Max. :12.000 전남 : 16561 Max. :2.000 Max. :12.000   
## (Other):102266   
## 사망시간 사망연령 사망장소   
## Min. : 0.00 Min. : 0.00 의료기관 :210292   
## 1st Qu.: 6.00 1st Qu.: 65.00 주택 : 43082   
## Median :11.00 Median : 77.00 사회복지시설: 11814   
## Mean :11.46 Mean : 73.44 기타 : 7693   
## 3rd Qu.:17.00 3rd Qu.: 85.00 병원 이송 중: 2893   
## Max. :99.00 Max. :999.00 도로 : 2015   
## (Other) : 3038   
## 사망자직업 혼인상태 교육정도   
## 학생, 가사, 무직 :203134 미혼 : 20824 불취학 :60370   
## 농림어업숙련종사자 : 27846 배우자있음:129172 초등학교 :89488   
## 서비스종사자및판매종사자: 10974 이혼 : 22314 중학교 :36904   
## 단순노무종사자 : 9681 사별 :108051 고등학교 :55086   
## 미상,군인(사병제외) : 7907 미상 : 466 대학교 :25229   
## 전문가및관련종사자 : 6544 대학원이상: 3759   
## (Other) : 14741 미상 : 9991   
## 사망원인   
## 악성신생물(Malignant neoplasms) :78194   
## 심장 질환(Heart diseases) :29735   
## 뇌혈관 질환(Cerebrovascular diseases) :23415   
## 폐렴(Pneumonia) :16476   
## 고의적자해(자살) (Intentional self-harm ):13092   
## (Other) :55484   
## NA's :64431

head(death)

## 신고날짜(월) 사망자주소 사망자성별 사망일자(월) 사망시간 사망연령  
## 1 1 인천 1 1 6 70  
## 2 1 경기 1 1 7 0  
## 3 1 서울 1 1 14 0  
## 4 1 서울 1 1 3 71  
## 5 1 서울 1 1 7 52  
## 6 1 서울 1 1 8 75  
## 사망장소 사망자직업 혼인상태 교육정도  
## 1 주택 학생, 가사, 무직 배우자있음 초등학교  
## 2 의료기관 학생, 가사, 무직 미혼 불취학  
## 3 의료기관 학생, 가사, 무직 미혼 불취학  
## 4 의료기관 학생, 가사, 무직 배우자있음 고등학교  
## 5 의료기관 학생, 가사, 무직 배우자있음 초등학교  
## 6 의료기관 학생, 가사, 무직 배우자있음 중학교  
## 사망원인  
## 1 뇌혈관 질환(Cerebrovascular diseases)  
## 2 출생전후기에기원한특정병태  
## 3 출생전후기에기원한특정병태  
## 4 악성신생물(Malignant neoplasms)  
## 5 간 질환(Diseases of liver)  
## 6 간 질환(Diseases of liver)

death<-death[death$사망연령 != 999,] #999세(미상) 제거  
  
suicide<-subset(death, 사망원인 == "고의적자해(자살) (Intentional self-harm )")  
summary(suicide) #요약

## 신고날짜(월) 사망자주소 사망자성별 사망일자(월)   
## Min. : 1.000 경기 :2879 Min. :1.000 Min. : 1.000   
## 1st Qu.: 4.000 서울 :2261 1st Qu.:1.000 1st Qu.: 4.000   
## Median : 7.000 부산 : 943 Median :1.000 Median : 6.000   
## Mean : 6.567 경남 : 912 Mean :1.294 Mean : 6.507   
## 3rd Qu.: 9.000 경북 : 773 3rd Qu.:2.000 3rd Qu.: 9.000   
## Max. :12.000 인천 : 769 Max. :2.000 Max. :12.000   
## (Other):4555   
## 사망시간 사망연령 사망장소   
## Min. : 0.00 Min. : 11.00 주택 :5638   
## 1st Qu.: 7.00 1st Qu.: 40.00 기타 :3526   
## Median :12.00 Median : 52.00 의료기관 :2360   
## Mean :11.72 Mean : 52.99 상업.서비스시설: 549   
## 3rd Qu.:17.00 3rd Qu.: 67.00 도로 : 312   
## Max. :23.00 Max. :109.00 병원 이송 중 : 268   
## (Other) : 439   
## 사망자직업 혼인상태 교육정도   
## 학생, 가사, 무직 :7383 미혼 :3776 불취학 : 762   
## 서비스종사자및판매종사자:1380 배우자있음:5803 초등학교 :2173   
## 단순노무종사자 : 824 이혼 :2063 중학교 :1773   
## 전문가및관련종사자 : 677 사별 :1408 고등학교 :4873   
## 농림어업숙련종사자 : 677 미상 : 42 대학교 :2761   
## 사무종사자 : 602 대학원이상: 262   
## (Other) :1549 미상 : 488   
## 사망원인   
## 고의적자해(자살) (Intentional self-harm ) :13092   
## 콜레라 (Cholera) : 0   
## 감염성및상세불명기원의기타위장염및대장염 : 0   
## 기타장감염성질환 (Other intestinal infectious disease ): 0   
## 호흡기 결핵(Respiratory tuberculosis) : 0   
## 기타 결핵(Other tuberculosis) : 0   
## (Other) : 0

###################################################################################################################  
  
# Step 1. 2016년 사망원인 데이터 읽어오기  
survey <- read.csv("자살충동여부조사.csv", header=FALSE)  
  
#분석에 사용할 항목만 데이터 선택하  
survey<- survey[,c(1:14, 18,21)]  
names(survey) <- c("성별", "만나이", "교육정도", "혼인상태", "흡연여부", "금연어려운이유",   
 "음주여부", "직장스트레스", "일상스트레스", "자살충동", "자살충동이유",  
 "교육비부담인식", "부모생존여부", "이혼에대한견해", "가족관계", "가구소득")  
  
#데이터 범위 설정  
survey <-subset(survey, 만나이 >= 50)  
survey <- subset(survey, 만나이 < 60)  
survey <- subset(survey, 자살충동 == 1)  
#성별 전처리  
survey$성별 <- factor(survey$성별, levels = c(1,2), labels=c("Male", "Female"))  
#교육정도 전처리  
survey$교육정도 <- factor(survey$교육정도, levels=c(0,1,2,3,4,5,6,7),  
 labels=c("lack of schooling", "elementary school", "middle school", "high school", "two-year univ.",   
 "four-year univ.", "postgraduate master's", "graduate doctor"))  
#혼인상태 전처리  
survey$혼인상태 <- factor(survey$혼인상태, levels=c(1,2,3,4), labels=c("single", "has a spouse",  
 "bereavement", "divorce"))  
#흡연여부 전처리  
survey$흡연여부 <- factor(survey$흡연여부, levels=c(1,2), labels=c("피운다", "피우지않는다"))  
#금연이어려운이유 전처리  
survey$금연어려운이유 <- factor(survey$금연어려운이유,  
 levels=c(1,2,3,4,5), labels=c("stress", "influence of others",  
 "withdrawal symptoms", "habbit", "etc"))  
#자살충동이유 전처리  
survey$자살충동이유 <- factor(survey$자살충동이유, levels=c(1,2,3,4,5,6,7,8,9),  
 labels=c("financial problem", "heterosexual problem", "physical.mental illness",  
 "work problem", "loneliness", "family discord", "school grade",  
 "being bullied", "etc"))  
#부모생존여부 전처리  
survey$부모생존여부 <- factor(survey$부모생존여부, levels=c(1,2,3,4),  
 labels=c("all of parents are alive", "only father is alive", "only mother is alive", "all of parents died"))  
#이혼에대한견해 전처리  
survey$이혼에대한견해 <- factor(survey$이혼에대한견해, levels=c(1,2,3,4,5),  
 labels=c("never", "not necessarily",  
 "may", "divorce if there is a reason", "not sure"))  
#가구소득 전처리 . 단위: 백만원  
survey$가구소득 <- factor(survey$가구소득, levels=c(1,2,3,4,5,6,7,8),  
 labels=c("~100", "100~200", "200~300", "300~400",  
 "400~500", "500~600", "600~700", "700~"))  
#교육비부담인식 전처리  
#survey$교육비부담인식 <- factor(survey$교육비부담인식, levels=c(1,2,3,4,5),  
# labels=c("매우부담", "약간부담", "보통", "별로부담X", "전혀부담X"))  
summary(survey)

## 성별 만나이 교육정도   
## Male :195 Min. :50.00 high school :215   
## Female:271 1st Qu.:52.00 middle school : 93   
## Median :55.00 elementary school : 58   
## Mean :54.63 four-year univ. : 45   
## 3rd Qu.:57.00 two-year univ. : 36   
## Max. :59.00 postgraduate master's: 11   
## (Other) : 8   
## 혼인상태 흡연여부 금연어려운이유  
## single : 31 피운다 :134 stress : 48   
## has a spouse:313 피우지않는다:332 influence of others: 2   
## bereavement : 31 withdrawal symptoms: 5   
## divorce : 91 habbit : 13   
## etc : 1   
## NA's :397   
##   
## 음주여부 직장스트레스 일상스트레스 자살충동  
## Min. :1.000 Min. :1.000 Min. :1.00 Min. :1   
## 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:1.00 1st Qu.:1   
## Median :1.000 Median :2.000 Median :2.00 Median :1   
## Mean :1.365 Mean :2.991 Mean :1.82 Mean :1   
## 3rd Qu.:2.000 3rd Qu.:5.000 3rd Qu.:2.00 3rd Qu.:1   
## Max. :2.000 Max. :5.000 Max. :4.00 Max. :1   
##   
## 자살충동이유 교육비부담인식   
## financial problem :202 Min. :1.000   
## physical.mental illness: 85 1st Qu.:1.000   
## family discord : 69 Median :1.000   
## loneliness : 68 Mean :1.735   
## work problem : 24 3rd Qu.:2.000   
## heterosexual problem : 11 Max. :5.000   
## (Other) : 7 NA's :368   
## 부모생존여부 이혼에대한견해  
## all of parents are alive: 51 never : 46   
## only father is alive : 17 not necessarily :129   
## only mother is alive :150 may :192   
## all of parents died :221 divorce if there is a reason: 83   
## NA's : 27 not sure : 16   
##   
##   
## 가족관계 가구소득   
## Min. :1.000 100~200:126   
## 1st Qu.:2.000 ~100 : 96   
## Median :3.000 200~300: 83   
## Mean :2.901 300~400: 63   
## 3rd Qu.:3.000 400~500: 41   
## Max. :6.000 500~600: 25   
## (Other): 32

#음주여부 1.있다 2. 없다  
#직장 및 일상 스트레스 1.매우 느낌 2. 느끼는 편 3. 느끼지 않는 편  
# 4. 전혀 느끼지 않음 5. 해당없음  
#형제, 부모, 가족관계 1. 매우만족, 2, 약간 만족, 3, 보통,  
# 4. 약간 불만족, 5.매우 불만족, 6. 해당없음